QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 8

1. Based on the Boston dataset available on the library MASS, relative to Housing Values in Suburbs of Boston. The variables of interest are:

   - Y equal to $medv$ is median value of owner-occupied homes in $\$1000$.
   - X_1 equal to $lstat$ is the lower status of the population (percent)
   - X_2 equal to rm is the average number of rooms per dwelling
   - X_3 equal to age is the proportion of owner-occupied units built prior to 1940

   (a) We fit the Model 1: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
   We use the following R commands for loading the data:

```
> library(MASS)
> data("Boston")
> attach(Boston)
The following objects are masked from Boston (pos = 13):

    age, black, chas, crim, dis, indus, lstat, medv, nox,
    ptratio, rad, rm, tax, zn
```

   Then we fit the Model 1 to the data:

```
> fitlm1 <- lm(medv ~ lstat + rm + age)
> summary(fitlm1)

Call:
lm(formula = medv ~ lstat + rm + age)

Residuals:
    Min      1Q  Median      3Q     Max
-18.210  -3.467  -1.053   1.957  27.500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.175311   3.181924  -0.369    0.712
lstat       -0.668513   0.054357 -12.298   <2e-16 ***
rm           5.019133   0.454306  11.048   <2e-16 ***
age          0.009091   0.011215   0.811    0.418
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.542 on 502 degrees of freedom
Multiple R-squared:  0.639,Adjusted R-squared:  0.6369
F-statistic: 296.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

(b) Looking at the last line of the command summary, we find that the F-Test is equal to 296.2 and there is strong evidence against the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and the R^2 is equal to 63% similar to the adjusted R^2.

(c) Moving to the parameters of interest, we look at the summary described above. In this case, we have that there is evidence to reject the null hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$ for β_1 and β_2, thus the coefficients for $lstat$ and rm are statistically significant. On the other hand, the intercept and the parameter related to age could not reject the null hypothesis, thus the two coefficients are not statistically significant.

(d) Moving to the second model, we fit the following Model 2: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. By using the R command, we have:

```
> fitlm2 <- lm(medv ~ lstat + rm)
> summary(fitlm2)

Call:
lm(formula = medv ~ lstat + rm)

Residuals:
    Min      1Q  Median      3Q     Max
-18.076  -3.516  -1.010   1.909  28.131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.35827    3.17283  -0.428    0.669
lstat       -0.64236    0.04373 -14.689   <2e-16 ***
rm           5.09479    0.44447  11.463   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.54 on 503 degrees of freedom
Multiple R-squared:  0.6386,Adjusted R-squared:  0.6371
F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

As previously described, we have that the F-statistic is 444.3, thus the overall regression is statistically significant and there is strong evidence against the null hypothesis. Moving to the parameters, in this scenario the parameter of $lstat$ and rm are statistically significant, while the intercept continuously remains not statistically significant.

(e) Regarding the best model, we compare the adjusted R^2 for both the models. For Model 1, $adj(R^2) = 0.6369$, while for Model 2, $adj(R^2) = 0.6371$, thus the Model 2 is the best model and in this case also all the parameters except the intercept are statistically significant.

2. **Coursework component**
When fitting the model
$$E[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$
to a set of $n = 25$ observations, the following results were obtained using the general linear model notation:
$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{pmatrix}, \qquad \boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix}$$
$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.11321519 & -0.00444859 & -0.000083673 \\ -0.00444859 & 0.00274378 & -0.000047857 \\ -0.00008367 & -0.00004786 & 0.000001229 \end{pmatrix}$$
Also $\boldsymbol{Y}^t\boldsymbol{Y} = 18310.63$ and $\bar{Y} = 22.384$.

(a) We find the least square estimator by using
$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \\ &= \begin{pmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{pmatrix}^{-1} \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix} \\ &= \begin{pmatrix} 2.34123 \\ 1.61591 \\ 0.01438 \end{pmatrix}\end{aligned}$$
Thus the related fitted model can be written as
$$y = 2.34123 + 1.61591x_1 + 0.01438x_2$$

(b) Based on the previous results, we can construct the ANOVA table. First of all, we need to define
$$SS_R = \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{Y} - n\bar{y}^2 = \begin{pmatrix} 2.34123 & 1.61591 & 0.01438 \end{pmatrix} \cdot \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix} - 25 \cdot 22.384^2$$
$$= 18076.9 - 12526.09 = 5550.81$$
Moving to the SS_T, we have that
$$SS_T = \boldsymbol{Y}^t\boldsymbol{Y} - n\bar{y}^2 = 18310.63 - 12526.09 = 5784.54$$
Thus, we have that $SS_E = SS_T - SS_R = 5784.54 - 5550.81 = 233.73$. Moving to S^2 or the so called MS_E, we have
$$S^2 = \frac{SS_E}{(25-3)} = \frac{233.73}{22} = 10.62409$$

Analogously, we have the MS_R, which is

$$MS_R = \frac{SS_R}{25-23} = \frac{5550.81}{2} = 2775.405$$

Finally the F statistic, which is

$$F = \frac{MS_R}{MS_E} = \frac{2775.405}{10.62409} = 261.237$$

In conclusion, we can build up the ANOVA table as

Source	df	SS	MS	F	p-value
Regression	2	5550.81	2775.405	261.237	0.0000
Residual	22	233.73	10.62409		
Total	24	5784.54			

3. Based on the previous results:
   (a) We see from the ANOVA table that the value of F is 261.237. The critical value at the 5% significance level for an F_{22}^{2} distribution is 3.44 so we can reject the null hypothesis at the 5% significance level.
   (b) Moving to the 95% confidence interval for the three parameters of the model. We start with β_0 and we remind that the variance of β_j is

$$S^2 \cdot c_{jj}$$

dove c_{jj} is the jth diagonal element of $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$. Remind in this case that $t_{22}(0.025) = 2.074$.

Thus, $\widehat{\beta}_0 = 2.34123$ and then its variance is $S^2 \cdot 0.11321519 = 10.62409 \cdot 0.11321519 = 1.20280$ and then its standard error is the square root of the variance, $\sqrt{1.20280} = 1.096722$.

Thus for β_0 the 95% confidence interval is

$$2.34123 \pm 2.074 \cdot 1.096722 = 2.34123 \pm 2.274601 = (0.066629, 4.615831)$$

Analogously, we can run the confidence interval for the other two coefficients β_1 and β_2. For β_1, we have that the standard error is equal to $\sqrt{S^2 \cdot 0.00274378} = \sqrt{0.02915017} = 0.1707342$. Thus, the 95% confidence interval for β_1 is

$$1.61591 \pm 2.074 \cdot 0.1707342 = 1.61591 \pm 0.3541027 = (1.261807, 1.970013)$$

In conclusion, for β_2, we have the following standard error $\sqrt{S^2 \cdot 0.000001229} = 0.003613448$ and thus the 95% confidence interval for β_2 is

$$0.01438 \pm 2.074 \cdot 0.003613448 = 0.01438 \pm 0.007494291 = (0.006885709, 0.02187429)$$

As one can see in every confidence interval, the zero is not included, thus all the coefficients are statistically significant at 5%.